The Challenge of *Deduplication* in Person-Centric Systems

Lessons from Immunization Registries and Integrated Child Health Information Systems (CHIS)

PHIN Conference
Atlanta GA May 24-27, 2004

All Kids Count

Public Health
INFORMATICS
Institute

Deduplication Technology and Practices for Integrated Child-Health Information Systems*

Susan M. Salkowitz, Consultant, Salkowitz Associates, LLC: salkowit@hln.com

Dr. Stephen Clyde, Computer Sciences Dep't, Utah State University: swc@cs.usu.edu

Ellen Wild, Director of Programs; All Kids Count, Public Health Informatics Institute: Ewild@taskforce.org

Preparation of this publication was supported by a contract from All Kids Count, a program of the Robert Wood Johnson Foundation

Objectives of Presentation

- Define the problem of finding and resolving duplicate records in person-centric information systems
- Describe the approaches that Immunization Registries and Integrated Child Health Systems (CHIS) have taken re: *deduplication*
- Provide an overview of the AKC *Connections* study ***Deduplication Technology and Practices for Integrated Child Health Information Systems***
- Demonstrate the utility of the study methodology and templates for PHIN
- Recommend some areas for Registry/CHIS/PHIN collaboration around *deduplication* protocols.

Deduplication -what is it?

- Integrated Child Health Information Systems (CHIS) are person-centric systems (often including Immunization Registries) which collect data from disparate files with different business rules for identification.
- This process can generate possible duplicate records.
- CHIS projects are challenged to resolve exact, near or alternate identity matches.
- CHIS use combinations of automated and manual methods for data cleaning activities termed *deduplication* to match and merge records appropriately and to prevent and remove duplicate records from the database.

Registry Standard Addresses *Deduplication*

- Immunization Registries are among the first public health systems to populate their databases from Vital Records and to exchange data on a real time basis with multiple levels of public health departments, private providers, community health clinics, hospitals and health plans.
- Registries recognized the problem of multiple records for the same individual and coined the term *deduplication* as a quality assurance process to resolve and remove potential duplicates from the database.
- The National Vaccine Advisory Committee (NVAC) endorsed Registry Functional Requirements contains:
- **Standard #12 :Promote accuracy and completeness of registry data**
- Definition The registry has developed and implemented a **data quality protocol** to combine all available information relating to a particular individual into a single, accurate immunization record.

Registry *Deduplication* Test Cases

- NIP has developed a toolkit to assist immunization registries in the evaluation of their deduplication algorithms.
  - The test data set consists of test cases that are fictitious, but representative of known duplicate record problems in real data, based on the information provided by various registry personnel.

- The evaluation tool application will calculate sensitivity and specificity values for the registry's algorithms based on the test results.
  - The sensitivity value measures how well the system performs at recognizing known duplicate records.
  - The specificity is the value that reflects how accurate the duplicate record detection is by measuring the rate at which non-duplicate records are misidentified.

Need for a *Deduplication* Study

- CHIS projects are challenged to select the most effective and least costly *deduplication* tools and strategies for their environments.
  - How do they know which tools to select?
  - What are other projects using?
  - How do the tools work?
  - How effective are they?
  - What do they cost?

Deduplication Software- What's out there?- the *Connections* study

- *All Kids Count Connections* Program* funded a *Deduplication* Domain Analysis
  - Performed at Utah State University Computer Science Department
  - Researched *deduplication* software and approaches
  - Performed a technical analysis and limited testing using the CDC test data set
  - Documented the findings in matrices showing effectiveness, underlying approach, cost and other factors.
  - Presented conclusions and recommendations

***All Kids Count Connections** at the *Public Health* **INFORMATICS** *Institute is a peer to peer learning network of 11 state and local health departments engaged in developing and implementing integrated information systems.*

Scope of *Connections* Study

- Collaborative of 8 of the *Connections* Child Health Integration Projects which include Immunization Registries [KS,ME,MO, NYC, OR (2) RI, UT]
- Development of questionnaire to identify products and practices used by Connections projects
- Research to identify technology and products that support deduplication in some way , from academic and commercial worlds-vendors/consultants
- Categorization of approaches:
  - By class of technical approach
  - By prerequisite enabling technology or file types
  - By effectiveness
  - By cost
  - By user types

Scope of Study (2)

- Perform *Off-line* analysis on software for which documentation was
- available
- Examine CDC *deduplication* test algorithm and specifications
- Perform *Benchmark* testing on one product for which software was available using CDC test cases
- Compile matrices of results
- Observations and recommendations
- Publication of Report

Section 2- Overview of *Deduplication* Technology - a Tutorial

- To make the *deduplication* process more tractable, researchers and software developers divide it into 3 sub-problems
    - Data-item transformation
    - Matching
    - Merging
- Solutions to deduplication problems vary
    - in underlying technology
    - in how they can hook into information systems
- Integration Classifications
    - Standalone
    - Software development kits
    - Server based systems

Section 3-Software Evaluation- Framework and Methodology

- Level 1- (*Off-line)* to be done on all products which can be described and analyzed from product specifications without access to the product itself.
  - Study identified 29 products: 8 were prioritized by participants for Level 1 Analysis
- Level 2-*(Benchmark*)testing of products against a known test data set- the CDC test data.
  - Provision of demo (incomplete) software, limitations on the number of records that can be tested and limited reporting of results.
  - Benchmark testing completed on only one product- leading more to “lessons learned” than a true evaluation

Section 3- Software Evaluation Factors

- Level 1- (*Off-line)- all products*
    - Platform
    - Processors
    - Dependency on environment
    - Types of databases they work on
    - Algorithms they are using
    - Matching and merging
    - Approach: machine learning, probabilistic, etc.
    - SDK- software development kits
    - Data transformations

Section 3-Software Evaluation Factors (2)

Level 2- (*Benchmark)*

Study identified evaluation criteria and some tips for users.

- Information on costs, set up, processing and other factors.
- Matching accuracy
- Success- false positives, false negatives
- Efficiency
- Processing time/database size
- Actual set up times
- Matching accuracy
- Records left for human review

Difficulties of benchmark testing due to lack of cooperation from vendors, inadequate [redacted]ntation and access to test beds.

Section 4-*Deduplication* software and approaches of 8 *Connections* projects

- Table of questionnaire results
- Detailed description of scope of projects and *deduplication* products and approaches used.
  - Level of automation
  - Degree of record matching
  - Source of information/effective data element for matching
  - Deployment timetables
- Highlighted key issues of organization, technology and participation in *community of practice* that affect success.

Section 5- General Observations

- Many factors (technical, political, and organizational), affect a project's ability to use *deduplication* processes effectively.
- One size does not fit all, and a combination of products and approaches need to be used because of
  - the quality variability of source systems
  - degree of automation for matching, verifying and merging
  - the intended uses of integrated information.

Observations- Record Matching

- Record matching products are extensive and cannot be individually evaluated or kept up to date.
- The study provides a framework for analysis
- There is inconclusive data to conclude whether a scoring or weighted,fuzzy comparison approach is better.
- An integrated system must be prepared to evaluate itself using test data that is representative of the conditions found in its real data.
- Most systems view Vital Records as the best source of name information, but no single program emerged as a single source of valid demographic information.
- Approaches for using field combinations were examined.

Observations-Deployment Options

- All projects indicate they have front end and back end processes and have developed tools to facilitate the merge process.
- There is a great underestimation of the time and effort to plan and execute *deduplication* processes.
- The number of stakeholders and the amount of control over implementation decisions and timing impacts deployment time.
- A master-client index approach is more heavily impacted by decisions of individual stakeholders than an incremental approach that applies *deduplication* to specific files but its functionality may be worth the effort.

Observations: Non Technical Determinants

- Scope and organization of the integration effort affects success- strategic planning and project organization within the DOH important
- Programmatic vs. technical control- programs may feel loss of control over their data
- Centralized vs. decentralized approach-operations become an "orphan" from funding support. Deduplication is a necessary function, but politically fragile
- Intended use of integrated data is a major determinant of its degree of completeness and accuracy

Observations- Non Technical Drivers for Success

Lessons for PHIN

- Immunization registry practices highlighted *deduplication* as a problem and a process- and are a foundational element of integrated systems.
- Electronic Vital Records systems are the authoritative source of DOB information and experiences in birth/death matching contribute to integration knowledge.
- Program or legislative mandates for integration, academic research and strategic planning initiatives also support more effective identification, development and use of *deduplication* methods and tools.
- *Community of Practice*, knowledge sharing and lessons learned contribute to success and visibility.

Uses of the report

- The full report with all of the matrices and tables is accessible via the Institute web site at www.phii.org
- This study was done within a *Community of Practice* as a demonstration of knowledge sharing to advance the principles of public health informatics.
- The Questionnaire can be adapted or used by projects to categorize their own approaches.
- The matrices of product characteristics and performance are time-perishable but the methodology can be applied to assess new products and protocols.
- The tutorial and the tables can help projects understand the choices and trade-offs as they select *deduplication* products and strategies.
- The observations on organizational and other non-technology-related factors can inform the PHIN process as more systems and programs are included in the PHIN architecture.

Areas for Registry/CHIS/PHIN Collaboration on *Deduplication*

- Utilize the expertise of Immunization Registries and CHIS on deduplication through the Public Health Informatics Institute and the American Immunization Registry Association as *communities of practice*
- Improve Testing and Assessment
  - Develop a more robust set of data-quality metrics- going beyond the CDC Deduplication Toolkit
  - Create a tool for generating data sets (instead of providing a fixed data set) that are representative of locale-specific data characteristics
  - Identify a more robust set of measurement tools
- Review testing strategies and methods to provide insight into managing testing activities

Areas for Registry/CHIS/PHIN Collaboration on Deduplication (2)

- Identify useful data elements and types of comparisons
- Examine the impact of Privacy Issues especially with regard to disclosure and consent of PHI
- Further study of Birth-Death matching as the *gold standard*
- Provide organizational support and technical assistance